

# A framework for probabilistic inferences from imperfect models

Meng Li and David B. Dunson

Department of Statistical Sciences, Duke University  
meng.li@stat.duke.edu dunson@duke.edu

November 7, 2016

## Abstract

The Bayesian paradigm provides a natural way to deal with uncertainty in model selection through assigning each model in a list of models under consideration a posterior probability, with these probabilities providing a basis for inferences or used as weights in model-averaged predictions. Unfortunately, this framework relies on the assumption that one of the models in the list is the true model. When this assumption is violated, the model that is closest in Kullback-Leibler divergence to the true model is often assigned probability converging to one asymptotically. However, when all the models are imperfect, interpretation of posterior model probabilities is unclear. We propose a new approach which relies on evaluating parametric Bayesian models relative to a nonparametric Bayesian reference using Kullback-Leibler divergence. This leads to a new notion of *absolute* posterior model probabilities, which can be used to assess the quality of imperfect models. Some properties of this framework are described. We consider an application to linear model selection against a Gaussian process reference, providing simple analytic forms for routine implementation. The framework is illustrated through simulations and applications.

**Key words:** Gaussian process; Gibbs posterior; Kullback-Leibler divergence; Model selection; M-open; Non-parametric Bayes; Posterior probabilities.

## 1 Introduction

Suppose we have a list of parametric models under consideration  $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_k\}$  for the observations  $y^{(n)} = \{y_1, \dots, y_n\} \in \mathcal{Y}^n$  with  $\mathcal{Y}$  the sample space. Each model  $\mathcal{M}_j$  has a corresponding likelihood  $p(\cdot | \theta_j, \mathcal{M}_j)$ , with  $\theta_j \in \Theta_j$  a finite dimensional parameter, and  $\pi(\cdot | \mathcal{M}_j)$  a prior density for  $\theta_j$ . Model  $\mathcal{M}_j$  has the marginal likelihood  $L_j(y^{(n)}) = \int p(y^{(n)} | \theta_j, \mathcal{M}_j) \pi(\theta_j | \mathcal{M}_j) d\theta_j$ , for  $j = 1, \dots, k$ . Assuming equal prior probabilities for each model to ease notation, the usual Bayesian paradigm assigns model  $\mathcal{M}_j$  posterior probability

$$\text{pr}(\mathcal{M}_j | y^{(n)}) = \pi_j = \frac{L_j(y^{(n)})}{\sum_{l=1}^k L_l(y^{(n)})}, \quad \text{for } j = 1, \dots, k, \quad (1)$$

so that the  $j$ th model receives a probability weight proportional to the marginal likelihood of the data under that model. Although these model probabilities are routinely used and touted as an appealing aspect of Bayesian approaches, they are calculated based on the flawed assumption that one of the models in the list of models under consideration is exactly true, known as the  $\mathcal{M}$ -closed case. As it is broadly accepted that parametric models are never exactly true, one may wonder about the philosophical meaning of a posterior model probability, and additionally question how these probabilities behave when the list  $\mathcal{M}$  does not contain the true data generating model, a case referred to as  $\mathcal{M}$ -open or  $\mathcal{M}$ -complete (Bernardo & Smith, 1994). In the  $\mathcal{M}$ -complete case, the true model is assumed to be known but possibly too complex, therefore we wish to select a model in  $\mathcal{M}$  due to its simplicity, interpretability or computational tractability.

Although most Bayesians are aware of this issue, the broad pragmatic view is that posterior model probabilities provide a useful basis for inferences and predictions under model uncertainty even in the  $\mathcal{M}$ -open and  $\mathcal{M}$ -complete case. A widely touted result in defending this viewpoint is that asymptotically for regular parametric models, the posterior probability on the model that is closest to the true data-generating model in Kullback-Leibler divergence converges to one. Although this seems somewhat reassuring, when the loss function differs in important ways from Kullback-Leibler divergence, model misspecification can lead to poor behavior theoretically and in practice (Jiang & Tanner, 2008; Owhadi et al., 2015).

There have been a number of attempts to accommodate the  $\mathcal{M}$ -open and  $\mathcal{M}$ -complete cases in the Bayesian literature. One of the most popular approaches is to formulate the model selection problem in a decision theoretic framework (Bernardo & Smith, 1994; Gutiérrez-Peña et al., 2009; Clyde & Iversen, 2013), selecting the model, or *action*, in  $\mathcal{M}$  that maximizes the expected utility. The expected utility can be approximated either via a cross-validation type of method (Clyde & Iversen, 2013) or using a nonparametric prior (Gutiérrez-Peña & Walker, 2005; Gutiérrez-Peña et al., 2009). Cross validation can be computationally intensive, and, more seriously, maximizing the expected utility produces a single optimal model or action without uncertainty quantification. A substantial advantage of Bayesian approaches is the ability to quantify uncertainty in model selection and propagate this uncertainty in conducting inferences and predictions.

In this article, we propose a simple and novel definition of *absolute* model probabilities, which do not involve other models in a list or assume that any of the parametric models under consideration exactly generated the observed data. The proposed notion of model probabilities has a connection to a range of concepts in the literature including Boltzmann (1878). To estimate these model probabilities, we require knowledge of the oracle model that generated the data. Using a nonparametric Bayes surrogate for the oracle, we provide methods for estimation and inference. For computational tractability, we focus primarily on comparing linear models using a Gaussian process surrogate, but the framework is broad.

## 2 Model Probabilities

### 2.1 Definition of model probabilities

Let  $\mathcal{N}^*$  be the oracle model which generated the data and  $f^*$  be the corresponding density function. Let  $\text{KL}(f, g) = \int f \log(f/g)$ . For any model  $\mathcal{M}_j$  with density  $f_j$ , we define the following absolute model probabilities:

$$\pi_j = \exp\{-n\text{KL}(f^*, f_j)\}, \quad j = 1, \dots, k, \quad (2)$$

which equals the exponentiated negative Kullback-Leibler divergence between  $\mathcal{M}_j$  and the oracle model. This definition is closely related to the notion of an *extent* of a distribution, which was introduced by Campbell (1966) using the exponentiated entropy, with the relative entropy Kullback-Leibler divergence as a special case. This notion of extent has been overlooked since its introduction, and has not been used outside of information theory. Under (2)  $\pi_j \in (0, 1)$  since  $\text{KL}(f^*, f_j)$  is always nonnegative. However, simply obeying this constraint does not make  $\pi_j$  interpretable as a model probability or useful as a basis of inference. Below we establish various probabilistic interpretations of (2).

### 2.2 Relationship to Boltzmann's formulation

As it is not rational to assign any parametric model  $\mathcal{M}_j$  a positive probability of being exactly the true data-generating model, we seek an alternative probabilistic interpretation of (2). We start by considering discrete sample spaces with  $\mathcal{Y} = \{\dagger_1, \dots, \dagger_m\}$ . The probability mass function  $f^*$  under the oracle model places probability  $a_l$  on element  $\dagger_l$ , for  $l = 1, \dots, m$ , while the probability mass function  $f_j$  under model  $\mathcal{M}_j$  places probabilities  $b_1, \dots, b_m$  on these elements. Under the oracle model, the expected number of occurrences of  $\dagger_l$  in  $n$  trials is  $na_l$ , for  $l = 1, \dots, m$ ; we refer to these values as the oracle frequencies. The probability of obtaining these frequencies from  $n$  independent observations from  $f_j$  is

$$\text{Multinomial}(n, na_1, \dots, na_m; f_j) = \binom{n}{na_1, \dots, na_m} b_1^{na_1} \dots b_m^{na_m}. \quad (3)$$

As commented by Akaike (1985), Boltzmann (1878) derived that the probability in (3) is asymptotically equal to  $\exp\{-n\text{KL}(f^*, f_j)\}$  up to a multiplicative constant. Since  $\text{KL}(f^*, f^*) = 0$ ,

$$\exp\{-n\text{KL}(f^*, f_j)\} = \frac{\exp\{-n\text{KL}(f^*, f_j)\}}{\exp\{-n\text{KL}(f^*, f^*)\}} \approx \frac{\text{Multinomial}(n, na_1, \dots, na_m; f_j)}{\text{Multinomial}(n, na_1, \dots, na_m; f^*)}, \quad (4)$$

as  $n \rightarrow \infty$ . The right hand side of (4) is interpretable as the likelihood of obtaining the oracle frequencies under model  $\mathcal{M}_j$  relative to the likelihood under the oracle model. As the multinomial likelihood of the oracle frequencies is maximized under the oracle model  $f^*$ , the right hand side of (4) is between zero and one, with the value moving closer to one as model  $\mathcal{M}_j$  improves relative to the oracle.

Although the Boltzmann (1987) probabilistic interpretation of (2) is specific to discrete distributions, the justification can be extended to continuous sample spaces  $\mathcal{Y}$  by first partitioning  $\mathcal{Y}$  into bins  $\mathcal{Y}_1, \dots, \mathcal{Y}_m$  having  $\mathcal{Y} = \bigcup_{l=1}^m \mathcal{Y}_l$  and  $\mathcal{Y}_l \cap \mathcal{Y}_{l'} = \emptyset$  for all  $l \neq l'$ . Then, letting  $m \rightarrow \infty$  with bin size  $|\mathcal{Y}_l| \rightarrow 0$ , one obtains a limiting form of (4). Therefore, we can generally use the exponentiated entropy between a candidate model and true model as a type of absolute probability weight on the candidate model for both discrete and continuous distributions.

### 2.3 Decision rules, model probabilities and p-values

The proposed model probabilities can also be obtained by an explicit decision rule in the setting of hypothetical repeated experiments. Suppose we have  $m$  repeated experiments ( $t = 1, 2, \dots, m$ ), where the observations  $y_t^{(n)} = \{y_{t1}, \dots, y_{tn}\}$  are drawn independently from  $f^*$  and the different experiments are independent. Define the likelihood ratio statistic for testing model  $\mathcal{M}_j$  against the oracle model using data from experiment  $t$  as:

$$T_{jt}^{(n)} = \frac{\prod_{i=1}^n f_j(y_{ti})}{\prod_{i=1}^n f^*(y_{ti})}.$$

The geometric mean of these likelihood ratio statistics across repeated experiments is  $R_{jm} = (\prod_{t=1}^m T_{jt}^{(n)})^{1/m}$ . As  $m$  increases,  $R_{jm} \rightarrow \exp\{-n\text{KL}(f^*, f_j)\}$  almost surely according to the strong law of large numbers, so for sufficiently large  $m$ ,  $R_{jm} < 1$  almost surely.

We define a random decision rule in which  $Z_{jm} = 1$  corresponds to choosing model  $\mathcal{M}_j$  based on the data from  $m$  replicated experiments, with  $Z_{jm} = 0$  otherwise. Choosing model  $\mathcal{M}_j$  is an absolute model selection decision about the merits of model  $\mathcal{M}_j$ . Based on data from  $m$  repeated experiments, as our decision rule we let

$$Z_{jm} \sim \text{Bernoulli}(R_{jm} \wedge 1),$$

where we take the minimum of  $R_{jm}$  and one to remove the possibility of  $R_{jm} > 1$  for finite  $m$ . This decision rule will tend to set  $Z_{jm} = 1$  with high probability if  $f_j$  provides an accurate approximation to  $f^*$ , with accuracy judged relative to the sample size  $n$ ; as sample size becomes larger it is appropriate to ask more of a parametric model. If on average across experiments the information in data having a sample size of  $n$  is sufficient to clearly distinguish the parametric and oracle model, then the decision rule will tend to set  $Z_{jm} = 0$  with high probability. In such a case, model  $\mathcal{M}_j$  would hopefully be assigned a small probability  $\pi_j$ , suggesting that we should continue our search for an adequate parametric model.

By increasing the number of replicated experiments  $m$  and using the geometric mean of the likelihood ratio test statistics, we remove sensitivity to variability across experiments. Let  $R_j$  and  $Z_j$  denote the random variables corresponding to  $R_{jm}$  and  $Z_{jm}$ , respectively. In the limit as the number of experiments increases  $m \rightarrow \infty$ , we obtain that

$$\text{pr}(Z_j = 1) = \exp\{-n\text{KL}(f^*, f_j)\} = \pi_j.$$

Hence, the absolute model probability  $\pi_j$  corresponds to the probability of selecting model  $\mathcal{M}_j$  based on a randomized decision rule that assesses whether the data in a sample size of  $n$  have sufficient information to distinguish the parametric model under consideration from the oracle.

Letting  $T_j^{(n)}$  denote the likelihood ratio test statistic based on a single experiment and  $t_j^{(n)}$  be the observed value of  $T_j^{(n)}$ , Bahadur (1967) shows that under certain regularity conditions  $\pi_j$  is asymptotically the p-value of the likelihood ratio test under the null hypothesis that the data are generated from model  $\mathcal{M}_j$ :

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(T_j^{(n)} < t_j^{(n)} \mid H_0) = -\text{KL}(f^*, f_j).$$

Hence, the absolute model probability  $\pi_j$  also has a frequentist testing interpretation.

## 2.4 Conditional model probabilities

The definition of  $\pi_j$  in (2) provides an absolute probability for a particular model. It is often useful to define conditional model probabilities relative to alternative models in a pre-specified list  $\mathcal{M}$ . We define the conditional probability for model  $\mathcal{M}_j$  as

$$\pi_{j|\mathcal{M}} = \frac{\pi_j}{\sum_{l=1}^k \pi_l} = \frac{\exp \{ -n\text{KL}(f^*, f_j) \}}{\sum_{l=1}^k \exp \{ -n\text{KL}(f^*, f_l) \}}, \quad (5)$$

which is simply the absolute probability for model  $\mathcal{M}_j$ , which can be viewed as a marginal probability, divided by the sum of the corresponding probabilities for each of the models in  $\mathcal{M}$ .

The probabilities in (5) can be used to compare alternative parametric models. Equation (5) has the same form as the famous Boltzmann-Gibbs weights in statistical mechanics with unit inverse temperature, where  $\text{KL}(f^*, f_j)$  is the energy of model  $j$ . By defining conditional model probabilities relative to other models in the list  $\mathcal{M}$ , we obtain a direct alternative to posterior model probabilities used in Bayesian inferences under the  $\mathcal{M}$ -closed framework. We will later show that the probabilities in (5) are asymptotically equivalent to usual posterior model probabilities if  $f^* = f_j$  for some  $j \in \{1, \dots, k\}$ , so that the oracle model exactly corresponds to one of the candidate parametric models. Although we view this assumption as unrealistic, this property is nonetheless reassuring.

## 3 D-Bayes inferences

### 3.1 Estimation of model probabilities

The model probabilities  $\pi_j$  and  $\pi_{j|\mathcal{M}}$  cannot be calculated directly, because the oracle model  $f^*$  is unknown and models in  $\mathcal{M}$  typically contain unknown parameters. To allow  $f^*$  to be unknown, we introduce a nonparametric reference model  $\mathcal{N}$ , which can be considered to be sufficiently flexible to accurately approximate the oracle, with accuracy improving with sample size. The nonparametric reference has density  $f_0$  and parameter  $\theta_0$ . The absolute and conditional model probabilities given the model list  $\mathcal{M}$  become

$$\pi_j = \exp \{ -n\widetilde{\text{KL}}(f_0, f_j) \}, \quad \pi_{j|\mathcal{M}} = \frac{\exp \{ -n\widetilde{\text{KL}}(f_0, f_j) \}}{\sum_{l=1}^k \exp \{ -n\widetilde{\text{KL}}(f_0, f_l) \}}, \quad (6)$$

where  $\widetilde{\text{KL}}(f_0, f_j)$  is an estimate of the Kullback-Leibler divergence between model  $\mathcal{M}_j$  and the reference model  $\mathcal{N}$ . We propose the following two estimators: a posterior mean estimator

$$\widetilde{\text{KL}}_1(f_0, f_j) = \int \int \text{KL}\{f_0(\cdot \mid \theta_0), f_j(\cdot \mid \theta_j)\} \pi(\theta_j \mid y^{(n)}) \pi(\theta_0 \mid y^{(n)}) d\theta_j d\theta_0, \quad (7)$$

and an estimator based on posterior predictive densities

$$\widetilde{\text{KL}}_2(f_0, f_j) = \text{KL}(\widehat{f}_0, \widehat{f}_j), \text{ where } \widehat{f}_j(\cdot) = \int f_j(\cdot \mid \theta_j) \pi(\theta_j \mid y^{(n)}) d\theta_j, \text{ and } j = 0, 1, \dots, k. \quad (8)$$

These two estimators address the uncertainty of parameters  $(\theta_j, \theta_0)$  differently: the posterior mean estimator uses the posterior mean of  $\text{KL}\{f_0(\cdot \mid \theta_0), f_j(\cdot \mid \theta_j)\}$ , while the posterior predictive estimator uses the Kullback-Leibler divergence between predictive densities of each model. As shown later in Section 4, the two estimators have the same asymptotic behavior and converge to the minimum Kullback-Leibler divergence to the oracle model among  $\theta_j \in \Theta_j$  under mild conditions. In practice, one can use whichever approximation is most convenient, or even rely on a mixture of (7) and (8).

We refer to the quantities in expression (6) as *D-probabilities*, as they provide a divergence-based alternative to usual Bayesian posterior model probabilities. D-probabilities have several advantages over usual posterior model probabilities. They avoid requiring that any parametric model is true, and provide an absolute measure of the quality of a specific model that is useful in assessing model adequacy and goodness-of-fit, issues that are notoriously poorly addressed in the Bayesian literature. Also, D-probabilities avoid some of the well-known pitfalls of posterior model probabilities, such as large sensitivity to the prior choice without an agreed upon method of default prior specification. The main challenges in the use of D-probabilities include the need to choose a nonparametric reference model, and develop accurate approximation algorithms. The nonparametric Bayes literature provides a rich menu of possibilities for  $\mathcal{N}$ , ranging from Dirichlet processes (Ferguson, 1973) to Gaussian processes (Rasmussen & Williams, 2006); for a review, refer to Hjort et al. (2010).

For reasons of simplicity in exposition and computational ease, we focus on normal linear models with a Gaussian process reference for the remainder of the article. In this case, conditional on covariance parameters, the Kullback-Leibler divergence between each parametric model and the nonparametric model can be calculated analytically, allowing us to rapidly conduct analyses and more easily study properties of the proposed model probabilities. There is a rich literature showing appealing properties of Gaussian process priors, such as rate adaptive behavior in nonparametric regression (van der Vaart & van Zanten, 2009).

### 3.2 D-Bayes inference for linear models

Let  $\{(x_i, y_i) : x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p, y_i \in \mathbb{R}\}_{i=1}^n$  be independent and identically distributed observations following the model

$$y \mid x \sim N\{\mu(x), \sigma^2\}, \quad (9)$$

where  $x$  is a  $p$ -dimensional predictor and  $y$  is a univariate response. Let  $Y = (y_1, \dots, y_n)$  and  $X = (x_1^T, \dots, x_n^T)^T$ . Letting  $j = 0$  index the reference model  $\mathcal{N}$  and  $j = 1, \dots, k$  index the parametric models, we let  $\mu_j(\cdot)$  and  $\sigma_j^2$  denote the mean function and variance, respectively, for model  $j$ . According to the chain rule of the Kullback-Leibler divergence, the proposed  $\widetilde{\text{KL}}_t(f_0, f_j)$  is equal to the Kullback-Leibler divergence between the conditional densities of  $y$  given  $x$  followed by an expectation with respect to the distribution of  $x$ . We use the empirical distribution of  $x$  for both  $\widetilde{\text{KL}}_1(f_0, f_j)$  and  $\widetilde{\text{KL}}_2(f_0, f_j)$ , which for example means that the term in (7) is calculated by  $\text{KL}\{f_0(\cdot \mid \theta_0), f_j(\cdot \mid \theta_j)\} = \sum_{i=1}^n \text{KL}\{f_0(\cdot \mid \theta_0, x_i), f_j(\cdot \mid \theta_j, x_i)\}/n$ .

We use a Gaussian process prior for  $\mu_0(\cdot)$ , with  $\mu_0(\cdot) \mid \sigma, \tau, \lambda \sim \text{GP}\{0, \sigma^2 k(\cdot, \cdot; \lambda, \tau)\}$  and covariance function

$$k(x_{i_1}, x_{i_2}; \lambda, \tau) = \tau^2 \exp \left\{ \sum_{j=1}^p \frac{-(x_{i_1,j} - x_{i_2,j})^2}{2\lambda_j^2} \right\}, \quad (10)$$

having predictor-specific bandwidth parameters  $\lambda = (\lambda_1, \dots, \lambda_p)^T$ . For notational convenience, we suppress the dependence of the covariance function on  $\lambda$  and  $\tau$ ; estimation of these hyper-parameters is discussed in Section 3.3. The prior distribution of  $\sigma_0^2$  is specified as  $p(\sigma_0^2) \propto 1/\sigma_0^2$ . Let  $K$  be the covariance matrix whose  $(i, j)$ th element is  $k(x_i, x_j)$ , and  $\mu_0^{(n)} = \{\mu_0(x_1), \dots, \mu_0(x_n)\}$  be the conditional mean vector. Then the reference model  $\mathcal{N}$  assumes  $Y \mid \mu_0^{(n)}, \sigma_0 \sim N\{\mu_0^{(n)}, \sigma_0^2 I_n\}$ , with priors  $\mu_0^{(n)} \mid \sigma_0 \sim N(0, \sigma_0^2 K)$  and  $p(\sigma_0^2) \propto 1/\sigma_0^2$ . Letting  $H = (I + K^{-1})^{-1} = K(K + I)^{-1}$ , we have

$$\mu_0^{(n)} \mid X, Y, \sigma_0 \sim N(HY, \sigma_0^2 H), \quad \sigma_0^2 \mid X, Y \sim \text{IG} \left\{ \frac{n}{2}, \frac{1}{2} Y^T (I - H) Y \right\}.$$

For model  $\mathcal{M}_j$ , let  $x_j$  be a  $p_j$ -dimensional sub-vector of  $x$  and  $\mu_j(x) = (1, x_j^T)\beta_j$ , so that model  $\mathcal{M}_j$  has parameters  $\theta_j = (\beta_j, \sigma_j^2)$ . Letting  $X_j$  denote the corresponding design matrix including a column of ones, the mean vector is  $\mu_j^{(n)} = \{\mu_j(x_1), \dots, \mu_j(x_n)\}^T = X_j\beta_j$ . With the following prior distributions:

$$\beta_j | \sigma_j^2 \sim N(0, \sigma_j^2 \Sigma_j), \quad p(\sigma_j^2) \propto 1/\sigma_j^2, \quad (11)$$

for some prior covariance matrix  $\Sigma_j$ , the posterior distributions are

$$\mu_j^{(n)} | \sigma_j^2, X_j, Y \sim N(H_j Y, \sigma_j^2 H_j), \quad \sigma_j^2 | X_j, Y \sim \text{IG}\left\{\frac{n}{2}, \frac{1}{2} Y^T (I - H_j) Y\right\},$$

where  $H_j = X_j(X_j^T X_j + \Sigma_j^{-1})^{-1} X_j^T$ .

The posterior mean estimates of  $\widetilde{\text{KL}}_1(f_0, f_j)$  in (7) are obtained as follows. Conditional on unknown parameters  $(\beta_j, \sigma_j, \mu_0^{(n)}, \sigma_0)$ , the Kullback-Leibler divergence between model  $\mathcal{M}_j$  and the reference model  $\mathcal{N}$  is

$$\text{KL}(f_0, f_j | \beta_j, \sigma_j, \mu_0^{(n)}, \sigma_0) = \frac{1}{2} \left\{ \frac{\sigma_0^2}{\sigma_j^2} + \frac{(X_j \beta_j - \mu_0^{(n)})^T (X_j \beta_j - \mu_0^{(n)})}{n \sigma_j^2} - 1 + \log \frac{\sigma_j^2}{\sigma_0^2} \right\}.$$

Using the fact that  $(X_j \beta_j - \mu_0^{(n)}) | \sigma_j, \sigma_0 \sim N\{(H_j - H)Y, \sigma_j^2 H_j + \sigma_0^2 H\}$ , we can further marginalize out  $\mu_0^{(n)}$  and  $\beta_j$  to obtain

$$\begin{aligned} \text{KL}(f_0, f_j | \sigma_j, \sigma_0) &= E_{\beta_j, \mu_0^{(n)} | \sigma_j, \sigma_0} \{ \text{KL}(f_0, f_j | \beta_j, \sigma_j, \mu, \sigma, \lambda, \tau) \} \\ &= \frac{1}{2} \left\{ \frac{\sigma_0^2}{\sigma_j^2} + \frac{Y^T (H_j - H)^2 Y + \text{tr}(\sigma_j^2 H_j + \sigma_0^2 H)}{n \sigma_j^2} - 1 + \log \frac{\sigma_j^2}{\sigma_0^2} \right\} \\ &= \frac{1}{2} \left\{ \frac{Y^T (H_j - H)^2 Y}{n \sigma_j^2} + \frac{\{1 + \text{tr}(H)/n\} \sigma_0^2}{\sigma_j^2} + \log \frac{\sigma_j^2}{\sigma_0^2} + \frac{\text{tr}(H_j)}{n} - 1 \right\}. \end{aligned}$$

By further integrating out  $\sigma_j$  and  $\sigma_0$ , we obtain that  $\widetilde{\text{KL}}_1(f_0, f_j) = E_{\sigma_0, \sigma_j} \text{KL}(f_0, f_j | \sigma_j, \sigma_0)$  as

$$\widetilde{\text{KL}}_1(f_0, f_j) = \frac{1}{n} (\mathcal{G}_{j,1} + \mathcal{P}_{j,1}), \quad (12)$$

where

$$\mathcal{G}_{j,1} = \frac{n}{2} \left[ \frac{Y^T (H_j - H)^2 Y}{Y^T (I - H_j) Y} + \frac{\{\text{tr}(H) + n\} Y^T (I - H) Y}{(n - 2) Y^T (I - H_j) Y} + \log \frac{Y^T (I - H_j) Y}{Y^T (I - H) Y} - 1 \right],$$

and  $\mathcal{P}_{j,1} = \text{tr}(H_j)/2$ .

We next obtain  $\widetilde{\text{KL}}_2(f_0, f_j)$ . Conditional on the variance parameters, the posterior predictive densities evaluated at  $\{x_1, \dots, x_n\}$  under the reference model and model  $\mathcal{M}_j$  are  $N\{HY, \sigma_0^2(I + H)\}$  and  $N\{H_j Y, \sigma_j^2(I + H_j)\}$ , respectively. Therefore, the Kullback-Leibler divergence  $\widetilde{\text{KL}}(f_0, f_j)$  conditional on the variances  $(\sigma_j^2, \sigma_0^2)$  is

$$\begin{aligned} \widetilde{\text{KL}}_2(f_0, f_j | \sigma_j, \sigma_0) &= \frac{1}{2} \left[ \frac{Y^T (H_j - H)^T (I + H_j)^{-1} (H_j - H) Y}{n \sigma_j^2} \right. \\ &\quad \left. + \frac{\sigma_0^2 \text{tr}\{(I + H_j)^{-1} (I + H)\}}{\sigma_j^2 n} + \log \frac{\sigma_j^2}{\sigma_0^2} + \frac{1}{n} \log \frac{\det(I + H_j)}{\det(I + H)} - 1 \right]. \end{aligned}$$

Integrating out the variance parameters  $\sigma_0^2$  and  $\sigma_j^2$  leads to

$$\widetilde{\text{KL}}_2(f_0, f_j) = \frac{1}{n} (\mathcal{G}_{j,2} + \mathcal{P}_{j,2}), \quad (13)$$

where

$$\begin{aligned}\mathcal{G}_{j,2} = & \frac{n}{2} \left[ \frac{Y^T(H_j - H)^T(I + H_j)^{-1}(H_j - H)Y}{Y^T(I - H_j)Y} + \frac{Y^T(I - H)Y}{Y^T(I - H_j)Y} \frac{\text{tr}\{(I + H_j)^{-1}(I + H)\}}{n - 2} \right. \\ & \left. + \log \frac{Y^T(I - H_j)Y}{Y^T(I - H)Y} - \log \det(I + H) - 1 \right],\end{aligned}$$

and  $\mathcal{P}_{j,2} = \log \det(I + H_j)/2$ .

Hence, both the posterior mean estimator in (12) and posterior predictive density estimator in (13) admit the decomposition of the form  $(\mathcal{G}_{j,t} + \mathcal{P}_{j,t})/n$  for  $t = 1, 2$ . Let the corresponding D-probabilities be  $\pi_{j,t} = \exp(-\mathcal{G}_{j,t} - \mathcal{P}_{j,t})$ . The term  $\mathcal{G}_{j,t}$  is the goodness-of-fit of model  $\mathcal{M}_j$  compared to the reference model and  $\mathcal{P}_{j,t}$  is a penalty term on model complexity. The trace of  $H_j$  is commonly used as the degrees of freedom of model  $\mathcal{M}_j$ , and the log determinant of the fitted covariance matrix  $\log \det(I + H_j)$  introduces a penalty on the rank of the covariance matrix (Fazel et al., 2003). Unlike most model selection criteria in the literature, the D-probability  $\pi_{j,t}$  is interpretable in an absolute sense for each candidate model, as discussed in Section 2.3. Therefore, the expression  $\mathcal{G}_{j,t}$  keeps any constant even when it is the same across all models.

If we use the flat prior where  $\Sigma_j^{-1} = 0$ , the matrix  $H_j$  is idempotent and we thus have  $\text{tr}(H_j) = p_j + 1$  and  $\log \det(I + H_j) = (p_j + 1) \log 2$ . Consequently, the D-probabilities penalize model complexity by

$$-\mathcal{P}_{j,1} = \frac{1}{2}(p_j + 1), \quad -\mathcal{P}_{j,2} = \frac{\log 2}{2}(p_j + 1). \quad (14)$$

When comparing two models  $\mathcal{M}_j$  and  $\mathcal{M}_{j'}$  where  $j \neq j'$ , the relative penalties on model complexity are the same as used in some existing criteria. Specifically, the penalty term  $\mathcal{P}_{j',1} - \mathcal{P}_{j,1} = (p_j - p_{j'})/2$  is used in the Akaike information criterions (Akaike, 1973, 1974) and the pseudo-Bayes factor (Geisser & Eddy, 1979), while  $\mathcal{P}_{j',2} - \mathcal{P}_{j,2} = \log 2(p_j - p_{j'})/2$  is the penalty term in the posterior Bayes factor (Aitkin, 1991; Gelfand & Dey, 1994).

### 3.3 Selection of hyperparameters

We estimate the parameters  $(\lambda, \tau)$  by maximizing the log marginal likelihood  $\log p(Y \mid \lambda, \tau)$ . Based on the log-likelihood of  $Y$  conditional on  $\{\mu_0^{(n)}, \sigma_0^2, \lambda, \tau\}$ , we first integrate out  $\mu_0^{(n)}$  to obtain

$$\begin{aligned}\log p(Y \mid \lambda, \tau, \sigma_0^2) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\sigma_0^2 K + \sigma_0^2 I| - \frac{1}{2} Y^T (\sigma_0^2 K + \sigma_0^2 I)^{-1} Y \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma_0^2 - \frac{1}{2} \log |K + I| - \frac{1}{2\sigma_0^2} Y^T (K + I)^{-1} Y,\end{aligned}$$

and further integrate out  $\sigma_0^2$ ,

$$\log p(Y \mid \lambda, \tau) = -\frac{1}{2} \log |K + I| - \frac{n}{2} \log \{Y^T (K + I)^{-1} Y\} + \text{constant}. \quad (15)$$

Let  $(\lambda_{\text{EB}}, \tau_{\text{EB}})$  be the empirical Bayes estimates maximizing equation (15). Then the D-probability of model  $\mathcal{M}_j$  is

$$\pi_j^{\text{EB}} = \exp\{-n\widetilde{\text{KL}}(f_0, f_j \mid \lambda_{\text{EB}}, \tau_{\text{EB}})\}.$$

To avoid conditioning on an empirical point estimate of  $(\lambda, \tau)$ , one may alternatively implement Markov chain Monte Carlo methods to draw posterior samples of the  $(p + 1)$ -dimensional parameter  $(\lambda, \tau)$  based on the likelihood in (15) and priors with positive supports such as gamma distributions. Let  $(\lambda^{(1)}, \tau^{(1)}), \dots, (\lambda^{(J)}, \tau^{(J)})$  be the posterior samples after burn-in, then the D-probability of model  $\mathcal{M}_j$  is

$$\pi_j^{\text{MCMC}} = \exp\left\{-\frac{n}{J} \sum_{i=1}^J \widetilde{\text{KL}}(f_0, f_j \mid \lambda^{(i)}, \tau^{(i)})\right\}.$$

## 4 Asymptotic behavior

In this section, we investigate the asymptotic behavior of the proposed  $\widetilde{\text{KL}}_t(f_0, f_j)$  in Sections 3 for linear models, and relate conditional D-probabilities with usual posterior model probabilities. We consider a compact support for the covariates, which is taken as  $[0, 1]^d$  without loss of generality. Let  $C[0, 1]^d$  be the space of continuous functions on  $[0, 1]^d$ . For a function  $g : [0, 1]^d \rightarrow \mathbb{R}$ ,  $x \in [0, 1]^d$  and  $i = 1, \dots, d$ , let  $g_i(\cdot | x)$  be a univariate function such that for any  $s \in [0, 1]$ ,  $g_i(s | x) = g(s')$  where  $s' = x$  but the  $i$ th element is replaced by  $s$ . Let  $\|g\|_\infty = \sup_{x \in [0, 1]^d} |g(x)|$  be the supremum norm of a function  $g$ , then the anisotropic Hölder space  $C^\alpha[0, 1]^d$  indexed by a vector of positive numbers  $\alpha = (\alpha_1, \dots, \alpha_d)$  contains all functions  $g$  such that for some  $L > 0$ ,

$$\sup_{x \in [0, 1]^d} \left\{ \sum_{j=0}^{\lfloor \alpha_i \rfloor} \|D^j g_i(\cdot | x)\|_\infty + \frac{\|D^{\lfloor \alpha_i \rfloor} g_i(y + h | x) - D^{\lfloor \alpha_i \rfloor} g_i(y | x)\|_\infty}{|h|^{\alpha_i - \lfloor \alpha_i \rfloor}} \right\} \leq L$$

for any  $(y, h)$  such that  $y \in [0, 1]$ ,  $h > 0$ ,  $y + h \in [0, 1]$  and  $i = 1, \dots, d$ ; here  $\lfloor \cdot \rfloor$  is the floor function and  $D^j$  is the  $j$ th derivative operator. Let  $\alpha_0^{-1} = \sum_{i=1}^d \alpha_i^{-1}$  be an exponent of global smoothness (Birgé, 1986; Barron et al., 1999; Hoffmann & Lepski, 2002).

For each model  $\mathcal{M}_j$ , we define

$$\theta_j^* = \arg \min_{\theta_j \in \Theta_j} \text{KL}\{f^*, f_j(\cdot | \theta_j)\}, \quad \delta_j = \text{KL}\{f^*, f_j(\cdot | \theta_j^*)\}.$$

The parameter value  $\theta_j^*$  is the so-called pseudotrue parameter (Bunke & Milhaud, 1998). A usual condition of Bayesian nonparametric models is that  $\delta_j = 0$  for all  $f^*$  in a large set of densities. Unless  $f^*$  exactly follows the parametric model under consideration, we have  $\delta_j > 0$  in general for any parametric model.

As the sample size  $n$  increases, the posterior measure for the density  $f$  under the nonparametric model  $\mathcal{N}$  will tend to concentrate in arbitrarily small Kullback-Leibler neighborhoods of the true data-generating model  $f^*$ . In contrast, the posterior measure for  $f$  under the parametric model  $\mathcal{M}_j$  will tend to concentrate on the point in the parametric class having the minimal Kullback-Leibler divergence from  $f^*$ . Heuristically, this type of behavior suggests that the proposed  $\widetilde{\text{KL}}_t(f_0, f_j)$  will tend to converge to the minimal Kullback-Leibler divergence within the support of model  $\mathcal{M}_j$  as  $n$  increases. However, as an information criterion, the Kullback-Leibler divergence may behave erratically (Barron, 1998), and the individual convergence of  $f_0$  and  $f_j$  does not directly imply the convergence of  $\widetilde{\text{KL}}_t(f_0, f_j)$  (Ikeda, 1960). We overcome these difficulties by taking advantage of the Gaussianity assumption on the errors, which allows us to relate the Kullback-Leibler divergence to well studied distances on model parameters. This is formalized in Theorem 4.1 based on the following assumptions.

- (a) Let  $\theta_0^* = \{\mu_0^*, \sigma_0^*\}$  be the true parameter values under model (9), which satisfy  $\mu_0^*(\cdot) \in C^\alpha[0, 1]^d$  and  $\sigma_0^* \in [a, b] \subset [0, \infty)$ . The covariate  $x$  is either fixed or randomly drawn from a density on  $[0, 1]^d$  that is bounded away from zero and infinity.
- (b) For the reference model, the regression function  $\mu_0$  has a squared exponential Gaussian process prior  $\Pi_\lambda$  as in (10); the prior distribution of  $\sigma_0$  has continuous density and is supported on  $[a, b]$ .
- (c) For each candidate model  $\mathcal{M}_j$ , the prior distribution of  $\theta_j = (\beta_j, \sigma_j)$  is supported on an open set  $\Theta_j \in \mathbb{R}^{p_j+1} \times \mathbb{R}^+$ , which has a continuous density that is bounded away from zero and infinity. The pseudotrue parameter  $\theta_j^*$  is unique and interior to  $\Theta_j$ .

**Theorem 4.1.** *Under model (9) and Assumptions (a), (b) and (c), for  $\widetilde{\text{KL}}_1(f_0, f_j)$  in (7), if  $\lambda_i = n^{\alpha_0/(2\alpha_0\alpha_i + \alpha_i)} (\log n)^{(d+1)\alpha_0/(2\alpha_0\alpha_i + \alpha_i)}$ , there exists a universal constant  $c > 0$  such that*

$$E_{\theta^*} |\widetilde{\text{KL}}_1(f_0, f_j) - \delta_j| \leq cn^{-\alpha_0/(2\alpha_0+1)} (\log n)^{(d+1)\alpha_0/(2\alpha_0+1)}.$$

For  $\widetilde{\text{KL}}_2(f_0, f_j)$  in (8), we have

$$0 \leq E_{\theta_0^*} \{\widetilde{\text{KL}}_2(f_0, f_j) - \delta_j\} \leq cn^{-\alpha_0/(2\alpha_0+1)} (\log n)^{(d+1)\alpha_0/(2\alpha_0+1)}$$



for sufficiently large  $n$ .

*Proof.* See the Appendix.  $\square$

**Remark 4.2.** The anisotropic Hölder space has been used to consider dimension-specific smoothness; for example, see Barron et al. (1999). The rate  $n^{-\alpha_0/(2\alpha_0+1)}$  is the minimax rate of convergence for a function in  $C^\alpha[0, 1]^d$  according to Hoffmann & Lepski (2002). We can obtain a rate-adaptive version of Theorem 4.1 without requiring the knowledge of  $\alpha$  to select  $\lambda$  by introducing an appropriate hyper-prior on  $\lambda$  following the random rescaling scheme in van der Vaart & van Zanten (2009) and Bhattacharya et al. (2014).

**Remark 4.3.** The Gaussian process prior in Assumption (b) can be replaced by any nonparametric priors that lead to nearly optimal contraction rate of the mean function under  $\|\cdot\|_n$  or a stronger metric, such as random series priors using a wavelet basis (Castillo, 2014) or B-splines (Yoo & Ghosal, 2016).

Theorem 4.1 suggests that the numerator in  $\pi_j$  given by (6) is approximately  $\exp(-n\delta_j)$  for large  $n$ . Consequently, for  $j_1 \neq j_2$ , the ratio  $\pi_{j_1}/\pi_{j_2}$  approximates  $\exp\{-n(\delta_{j_1} - \delta_{j_2})\}$ . The commonly used Bayes factor between two candidate models has been proven to have the same asymptotic behavior. For any two different models  $\mathcal{M}_{j_1}$  and  $\mathcal{M}_{j_2}$ , the Bayes factor  $\text{BF}(\mathcal{M}_{j_1}, \mathcal{M}_{j_2}) = I_n^{(j_1)}/I_n^{(j_2)}$  is approximately equal to  $\exp\{-n(\delta_{j_1} - \delta_{j_2})\}$  under mild conditions (Walker et al., 2004, Theorem 1), suggesting its asymptotic equivalence to the proposed D-probabilities. However, unlike Bayes factors which do not allow improper priors on model-specific parameters, D-probabilities are well defined under improper priors as long as the posteriors under each model are proper.

## 5 Simulation

In this section, we conduct simulations to investigate the finite sample performance of the proposed D-probabilities, while comparing with usual Bayesian approaches in various settings. We focus initial on a univariate case; Section 6 illustrates comparisons for multivariate cases. Under model (9), we generate the covariate  $x$  from the uniform distribution on  $(0, 1)$ , use  $\sigma = 1$  for the noise standard deviation, and let the sample size  $n = 100$ . We consider the model list  $\mathcal{M} = \{\mathcal{M}_F, \mathcal{M}_N\}$ , where the full model  $\mathcal{M}_F$  is the simple linear regression model and the null model  $\mathcal{M}_N$  only has the intercept. We index the two models by  $j = F$  and  $j = N$ . Throughout this section, we use the default prior in (11) with prior precision  $\Sigma_j^{-1} = 0$  for the parameters in model  $\mathcal{M}_j$  to calculate all D-probabilities. The number of replications is 1000.

We first consider the mean function

$$\mu(x) = 10 + \beta(\gamma)x + \gamma \log x, \quad (16)$$

where  $\beta(\gamma) = \{12(e^{1/10} - 1 - \gamma^2/4)\}^{1/2} - 3\gamma$  and  $\gamma$  is some positive constant in  $\Gamma = [0, 2(e^{1/10} - 1)^{1/2}]$ . According to Lemma 3 in the supplementary material, we can obtain that  $\delta_F = \{\log(1 + \gamma^2/4)\}/2$ , and the specification of  $\beta(\gamma)$  ensures that  $\delta_N = 0.05$  for all  $\gamma \in \Gamma$ . Therefore, the parameter  $\gamma$  controls how the mean function deviates from a linear model. We have the  $\mathcal{M}$ -closed situation when  $\gamma = 0$ , and  $\mathcal{M}$ -complete situation when  $\gamma > 0$ .

In addition to the D-probabilities, we estimate usual Bayesian model probabilities using Zellner  $g$  priors for the regression coefficients, with covariance  $\Sigma_j^{-1} = (X_j^T X_j)/g$ . We consider two choices of  $g$ : the unit information prior in which  $g = n$  (Kass & Wasserman, 1995), which leads to the Bayesian information criteria for model selection under some conditions, and the hyper- $g$  prior (Liang et al., 2008), which lets  $g/(g+1) \sim \text{Beta}(1, 1/2)$ . Both these priors have been implemented in the R package BAS.

We use 20 equal-spaced grid points from 0 to  $2(e^{1/10} - 1)^{1/2}$  for  $\gamma$ . Figure 1 plots various estimates versus  $\delta_F$ . Figure 1 (a) shows that the conditional D-probabilities  $\pi_{F,1|\mathcal{M}}$  are between the model probabilities under unit information and hyper- $g$  priors, while the alternative form  $\pi_{F,2|\mathcal{M}}$  tends to give larger D-probabilities due to the smaller penalty on model complexity as in (14). Figure 1 (b) presents the inclusion probability of the covariate  $x$ . The unit information prior and hyper- $g$  prior are observed to give smaller inclusion probability of

$x$  when  $\delta_F = \delta_N = 0.05$ , compared to D-probabilities. In this case, the mean function is  $\mu(x) = 10 - 1.95x + 0.65 \log x$ . The covariate  $x$  clearly impacts  $\mu(x)$  but all model probabilities tend to prefer the null model.

We next compared out-of-sample prediction accuracy based on the root mean squared error:  $\{\sum_{i \in \mathcal{T}} (\hat{Y}_i - Y_i)^2 / 100\}^{1/2}$  where  $\mathcal{T}$  is a validation set. For each method, we calculate the predictive mean under the highest probability model. As shown in Figure 1 (c), all methods have similar predictive performance. A small absolute D-probability, or equivalently a large  $\min(\delta_F, \delta_N)$ , suggests that the candidate model has poor fit relative to the nonparametric model  $f_0$ . One may conclude that none of the models in  $\mathcal{M}$  fit sufficiently well if they all have Kullback-Leibler divergence larger than a pre-specified threshold, such as  $-\log(c_\pi)/n$ , with  $c_\pi = 0.01$  implying a threshold of 0.046 on Kullback-Leibler divergence.

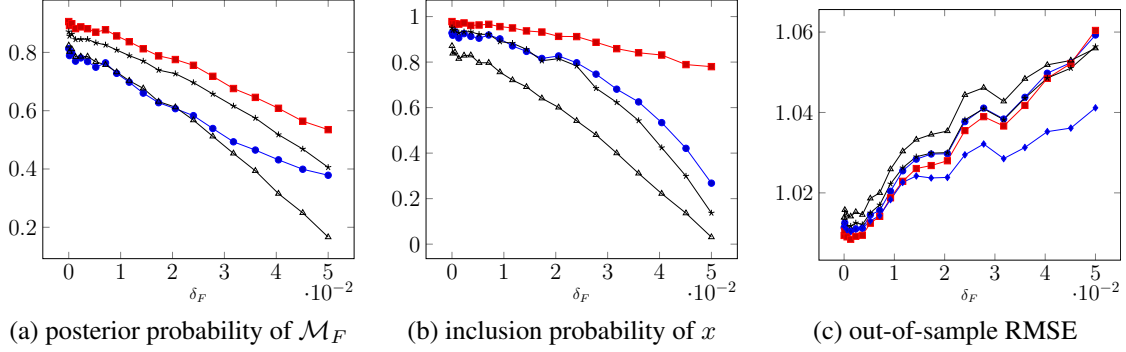


Figure 1: Comparison of D-probabilities versus other posterior model probabilities under model (16): conditional D-probabilities  $\pi_{F,1|\mathcal{M}}$  (circle), conditional D-probabilities  $\pi_{F,2|\mathcal{M}}$  (square), unit information prior (triangle) and hyper  $g$  prior (star). Plot (c) presents the out-of-sample root mean squared error or RMSE of the highest probability models selected by each method including the reference model (diamond). Results are based on 1000 replications.

We next consider another four cases with different mean functions:  $\mu_1(x) = 10 + 10x$ ,  $\mu_2(x) = 10$ ,  $\mu_3(x) = 10 + \sin(30\pi x)$  and  $\mu_4(x) = 10x^5$  where Case  $i$  uses the mean function  $\mu_i(x)$  for  $i = 1, 2, 3, 4$ . Case 1 and Case 2 are for the  $\mathcal{M}$ -closed situation where the model list  $\mathcal{M}$  contains the true model, Case 3 and Case 4 are for the  $\mathcal{M}$ -complete situation while Case 3 is close to an  $\mathcal{M}$ -open situation as the reference model is expected to fail to detect the high frequency oscillation. We vary the sample size  $n = (100, 500)$  and use 1000 replications. While detailed descriptions of this simulation are deferred to the supplementary material, we observe that both  $\widetilde{\text{KL}}_t(f_0, f_j)$  quickly converge to the corresponding  $\delta_j$  in Case 1, 2 and 4. Case 3 corresponds to a subtle cyclic deviation from Case 2; we find in this case that the reference nonparametric model fails to pick up the cyclic deviation so that the estimates of  $\widetilde{\text{KL}}_t(f_0, f_j)$  are close to Case 2 but deviate from  $\delta_j$ . However, the estimates of  $\delta_N - \delta_F$  are accurate, suggesting robustness of the conditional D-probabilities to performance of the reference model. In Case 1, the D-probability is higher for the true model  $\mathcal{M}_F$  in all replications, suggesting model selection uncertainty close to zero. Model  $\mathcal{M}_F$  has absolute D-probabilities that are not close to zero, such as 0.3, providing evidence it is an adequate approximation. In Case 2, both models have high D-probabilities as expected. The inclusion probability of the covariate  $x$  is either 0.09 using  $\pi_{j,1}$  or 0.23 using  $\pi_{j,2}$ , suggesting preference for the null model, with  $\pi_{j,1}$  providing a greater penalty on model complexity. The slight difference in scales between  $\pi_{j,1}$  and  $\pi_{j,2}$  is observed to be less prominent for conditional D-probabilities. In Case 4, the full model  $\mathcal{M}_F$  is assigned probability 1, but the D-probabilities are both close to zero, suggesting lack of fit.

## 6 Data application: ozone data

As an another illustration of the differences between our proposed D-probability based approach and usual Bayesian approaches to variable selection, we focus on ground-level ozone data (Breiman & Friedman, 1985;

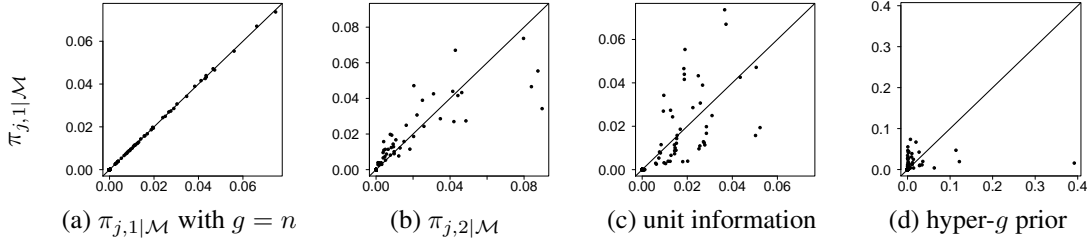


Figure 2: Comparison of  $\pi_{j,1|\mathcal{M}}$  versus other posterior model probabilities:  $\pi_{j,1|\mathcal{M}}$  with  $g = n$ ,  $\pi_{j,2|\mathcal{M}}$ , unit information and hyper- $g$  prior, from left to right.

Casella & Moreno, 2006; Liang et al., 2008). The ozone data, which are available in the R package *faraway*, consist of  $n = 330$  daily ozone readings in Los Angeles along with eight meteorological explanatory variables. We rescale each of these explanatory variables  $x$  to  $[0, 1]$  via the transformation  $(x - x_{\min}) / (x_{\max} - x_{\min})$  where  $x_{\max}$  and  $x_{\min}$  are the observed maximum and minimum values of  $x$ , respectively. The description of all variables are given in the supplementary material. The model list  $\mathcal{M}$  includes  $2^8 = 256$  candidate models corresponding to all possible subsets of explanatory variables.

We first calculate both versions of our D-probabilities,  $(\pi_{j,1}, \pi_{j,2})$ , for each candidate model following Section 3.2. Relative to usual Bayesian model probabilities, one of the appealing aspects of D-probabilities is the reduced sensitivity to the choice of prior distribution for model-specific parameters. In fact, we can even use default non-informative priors without the usual pitfalls. To illustrate this, we first considered the default prior in (11) with prior precision  $\Sigma_j^{-1} = 0$  for the parameters in model  $\mathcal{M}_j$ . In the Bayesian literature on variable selection in linear models, the most broadly used priors for the regression coefficients fall in the Zellner  $g$  family, and we consider the unit information prior and hyper- $g$  prior as in Section 5.

Figure 2 plots our conditional D-probabilities  $\pi_{j,1|\mathcal{M}}$  for each candidate model under a default prior against other choices, including (a)  $\pi_{j,1|\mathcal{M}}$  under a unit information prior, (b) the alternative form for the D-probabilities  $\pi_{j,2|\mathcal{M}}$  under a default prior, (c) usual Bayes model probabilities under a unit information prior, and (d) usual Bayes model probabilities under a hyper- $g$  prior. As expected, we found that D-probabilities were insensitive to slight changes in the prior distribution for the regression coefficients, with the values under the default prior essentially identical to those under a unit information prior. In addition, the two version of conditional D-probabilities were highly correlated. We also found that the D-probabilities were correlated and had similar magnitudes to the usual Bayes model probabilities under a unit-information prior, but differed dramatically from the Bayes model probabilities under a hyper- $g$  prior. In particular, the highest Bayes model probabilities under the hyper- $g$  prior were much larger than the highest D-probabilities. This is also illustrated in Table 1, which presents the model having the highest probability under each of the approaches. The top models based on  $\pi_{j,1|\mathcal{M}}$  and  $\pi_{j,2|\mathcal{M}}$  with default priors had probabilities 0.07 and 0.09, respectively. In contrast, the model having the highest usual Bayesian probability under the hyper- $g$  prior was 0.39, compared to a value of only 0.05 under a unit information prior. This serves in part to illustrate again the well known sensitivity of usual Bayesian model probabilities to the prior on the regression coefficients. Each of the four different approaches considered in the Table yielded somewhat different top models. This difference in ordering of top models is not unexpected given that the sample size is only  $n = 165$ , leaving out half the data to allow cross validation, and there are 256 models under consideration. To gauge the extent to which the data can distinguish between these different top models, we compared out-of-sample prediction accuracy based on the root mean squared error as in Section 5. As shown in the last column of Table 1, all of the models had essentially identical predictive performance. This is consistent with our expectation that the data are not sufficient to select from among a moderate number of top models, suggesting model probabilities in the single digits are more realistic than the 0.39 value produced by the hyper- $g$  prior.

Another unique aspect of the D-probability approach is the ability to provide absolute model probabilities instead of just values conditionally on falling in the list of possible linear models. We find in the ozone application that the absolute D-probabilities are extremely small for all of the candidate models, having a

Table 1: Selected variables and the corresponding posterior model probability using various methods on the entire dataset. The last column presents the out-of-sample root mean squared error or RMSE of the highest probability models selected by each method; results are based on 100 replications and the maximum standard errors is 0.02

Method	Variables in the model	Probability	RMSE
$\pi_{j,1 \mathcal{M}}$	vh, humidity, temp, ibh, ibt, vis	0.07	4.61
$\pi_{j,2 \mathcal{M}}$	vh, wind, humidity, temp, ibh, dpq, ibt, vis	0.09	4.61
unit information	humidity, temp, ibh, vis	0.05	4.62
hyper- $g$ prior	humidity, temp, ibh	0.39	4.63

maximum value of only  $1.65 \times 10^{-22}$ . This suggests that linear models provide a poor fit to the data relative to a nonparametric model; indeed, the root mean square error out of sample for the nonparametric reference model was significantly reduced to 4.09 from a minimum value of 4.61 for any of the linear models. Adding quadratic and interaction terms to expand the set of linear models leads to reductions to a range of 4.4 to 4.6 for root mean square errors out of sample (Liang et al., 2008), but there was still a significant gap in performance relative to the reference nonparametric model. This application has illustrated the practical advantages of D-probabilities relative to usual Bayes model probabilities in terms of reducing sensitivity to the prior and allowing the use of reference priors, while providing evidence of lack of fit of parametric models and producing a nonparametric reference as an alternative.

## Supplementary Material

Supplementary material includes two lemmas used in the proof of Theorem 4.1, a third lemma to gives analytical forms of the divergence  $\delta_j$  for linear regression models when the true model is (9), additional simulation results and the description of variables in the Ozone data.

## Appendix

### Proof of Theorem 4.1

Throughout this proof, we use the notation  $a \lesssim b$  if  $a \leq Cb$  for a universal constant, and  $a \asymp b$  if  $a \lesssim b \lesssim a$ . We first view the covariates  $x_1, \dots, x_n$  as fixed; for a function  $g : [0, 1]^d \rightarrow \mathbb{R}$ , we define the empirical norm  $\|g\|_n$  by  $\|g\|_n = \{\sum_{i=1}^n g^2(x_i)/n\}^{1/2}$ .

For the reference model  $\mathcal{M}_0$ , let  $\mathbb{H}^\lambda$  be the reproducing kernel Hilbert space of the Gaussian process prior  $\Pi_\lambda$  and  $\|\cdot\|_{\mathbb{H}^\lambda}$  be the associated norm; see Van Der Vaart & van Zanten (2008) for more technical details about the reproducing kernel Hilbert space. The so-called concentration function  $\phi_{\mu_0^*}$  is

$$\phi_{\mu_0^*}(\epsilon) = \inf_{h \in \mathbb{H}^\lambda : \|h - \mu_0^*\|_\infty < \epsilon} \|h\|_{\mathbb{H}^\lambda}^2 - \log P(\mu : \|\mu\|_\infty < \epsilon).$$

Letting  $\epsilon_n = \sup\{\epsilon > 0 : \phi_{\mu_0^*}(\epsilon) \geq n\epsilon^2\}$ , we have

$$E_{\theta_0^*} \int \|\mu_0 - \mu_0^*\|_n^2 \pi(\mu_0 | X, Y) d\mu_0 \lesssim \epsilon_n^2 \quad (17)$$

uniformly in the design points, in view of Theorem 1 in Van Der Vaart & Van Zanten (2011).

We next calculate the contraction rate  $\epsilon_n$  under the assumption that  $\mu_0^* \in C^\alpha[0, 1]^d$ . According to Lemma 4.2 and 4.3 in Bhattacharya et al. (2014), there exists constants  $C_1$  and  $C_2$  depending only on  $\mu_0^*$  and a constant  $C_3$  such that

$$\inf \left\{ \|h\|_{\mathbb{H}^\lambda}^2 : \|h - \mu_0^*\|_\infty \leq C_1 \sum_{i=1}^d \lambda_i^{-\alpha_i} \right\} \leq C_2 \prod_{i=1}^d \lambda_i,$$

$$-\log \Pi_\lambda(\mu : \|\mu\|_\infty \leq \epsilon) \leq C_3 \prod_{i=1}^d \lambda_i \left\{ \frac{\max(\lambda)}{\epsilon} \right\}^{d+1}.$$

For a sequence  $\epsilon_n \rightarrow 0$  and  $n\epsilon_n^2 \rightarrow \infty$ , we equate  $\epsilon_n \asymp \sum_{i=1}^d \lambda_i^{-\alpha_i}$  and  $\prod_{i=1}^d \lambda_i (\log n)^{d+1} \asymp n\epsilon_n^2$ . Letting  $\lambda_i = \epsilon_n^{-1/\alpha_i}$ , we then have  $n\epsilon_n^2 \asymp \epsilon_n^{-\alpha_0^{-1}} (\log n)^{d+1}$ . Consequently, the optimal choice of  $\lambda$  and the corresponding contraction rate are

$$\lambda_i = n^{\alpha_0/(2\alpha_0\alpha_i + \alpha_i)} (\log n)^{(d+1)\alpha_0/(2\alpha_0\alpha_i + \alpha_i)}, \quad \epsilon_n = n^{-\alpha_0/(2\alpha_0+1)} (\log n)^{(d+1)\alpha_0/(2\alpha_0+1)}.$$

In addition, the standard deviation  $\sigma_0$  has the same contraction rate  $\epsilon_n$ , that is

$$E_{\theta_0^*} \int |\sigma_0 - \sigma_0^*| \pi(\sigma_0 | X, Y) \lesssim \epsilon_n, \quad (18)$$

according to Theorem 3.3 in van der Vaart & van Zanten (2008).

For the candidate model  $\mathcal{M}_j$ , we shall apply the Bernstein-von Mises theorem under misspecification (Bunke & Milhaud, 1998; Kleijn & van der Vaart, 2012). Model  $\mathcal{M}_j$  is a finite-dimensional model with Gaussian noise and the true regression model has a smooth mean function, thus regularity conditions for asymptotic normality are satisfied; for example, see Remark 6 in Bunke & Milhaud (1998). Since  $\|\mu_j - \mu_j^*\|_n^2 = (X_j \beta_j - X_j \beta_j^*)^T (X_j \beta_j - X_j \beta_j^*) = (\beta_j - \beta_j^*)^T X_j^T X_j (\beta_j - \beta_j^*)$ , we have

$$E_{\theta_0^*} \int \|\mu_j - \mu_j^*\|_n^2 \pi(\beta_j | X, Y) d\beta_j \lesssim n^{-1}, \quad (19)$$

and

$$E_{\theta_0^*} \int |\sigma_j - \sigma_j^*|^2 \pi(\sigma_j | X, Y) d\sigma_j \lesssim n^{-1}, \quad (20)$$

uniformly in the design points  $(x_1, \dots, x_n)$ .

Let

$$\widetilde{\text{KL}}_1\{f_0(\cdot | \theta_0), f_j(\cdot | \theta_j)\} = \log \frac{\sigma_j}{\sigma_0} + \frac{\sigma_0^2 + \|\mu_j - \mu_0\|_2^2}{2\sigma_j^2} - \frac{1}{2},$$

and

$$\delta_j^{(n)} = \log \frac{\sigma_j^*}{\sigma_0^*} + \frac{\sigma_0^{*2} + \|\mu_j^* - \mu_0^*\|_2^2}{2\sigma_j^{*2}} - \frac{1}{2}.$$

Using the facts that both  $\sigma_j$  and  $\sigma_0$  have bounded supports and  $\widetilde{\text{KL}}_1\{f_0(\cdot | \theta_0), f_j(\cdot | \theta_j)\}$  is a continuous function of  $(\sigma_0, \sigma_j)$ , it is easy to verify that

$$\begin{aligned} |\widetilde{\text{KL}}_1(f_0, f_j) - \delta_j^{(n)}| &= |E[\widetilde{\text{KL}}_1\{f_0(\cdot | \theta_0), f_j(\cdot | \theta_j)\}] - \delta_j^{(n)}| \\ &\lesssim E|\sigma_j - \sigma_j^*| + E|\sigma_0 - \sigma_0^*| + |E\|\mu_j - \mu_0\|_2^2 - \|\mu_j^* - \mu_0^*\|_2^2| \\ &\lesssim E|\sigma_j - \sigma_j^*| + E|\sigma_0 - \sigma_0^*| + E\|\mu_j - \mu_j^*\|_2^2 + E\|\mu_0 - \mu_0^*\|_2^2, \end{aligned}$$

where the expectation  $E$  is taken with respect to the posterior distributions of the corresponding parameters. Combining equations (17), (18), (19) and (20), we obtain that

$$E_{\theta_0^*} |\widetilde{\text{KL}}_1(f_0, f_j) - \delta_j^{(n)}| \lesssim n^{-1/2} + \epsilon_n + n^{-1} + \epsilon_n^2 \leq c\epsilon_n,$$

for some universal constant  $c$  uniformly in the design points  $\{x_1, \dots, x_n\}$ .

For random designs where  $x \sim F$ , we have

$$\delta_j = \log \frac{\sigma_j^*}{\sigma_0^*} + \frac{\sigma_0^{*2} + \int |\mu_j^*(x) - \mu_0(x)|^2 dF}{2\sigma_j^{*2}} - \frac{1}{2}.$$

By a direct application of the central limit theorem and boundedness of  $\|\mu_j^* - \mu_0^*\|$ , we obtain  $E_F|\delta_j^{(n)} - \delta_j| \lesssim n^{-1/2}$ . Since  $\epsilon_n \gtrsim n^{-1/2}$ , the contraction rate of  $\widetilde{\text{KL}}_1$  is still  $\epsilon_n$ .

For  $\widetilde{\text{KL}}_2(f_0, f_j)$ , in view of Lemma 1 in the supplementary material, we have  $\widetilde{\text{KL}}_2(f_0, f_j) - \delta_j \leq \widetilde{\text{KL}}_1(f_0, f_j) - \delta_j$  thus  $E_{\theta_0^*}\{\widetilde{\text{KL}}_2(f_0, f_j) - \delta_j\} \leq c\epsilon_n$ . Equations (17), (18), (19) and (20) imply that  $\hat{f}_0 \rightarrow f_0^*$  and  $\hat{f}_j \rightarrow f_j^*$ , therefore  $\liminf_{n \rightarrow \infty} E_{\theta_0^*}\widetilde{\text{KL}}_2(f_0, f_j) \geq \delta_j$  according to Lemma 2 in the supplementary material. It follows that  $0 \leq E_{\theta_0^*}\widetilde{\text{KL}}_2(f_0, f_j) - \delta_j \leq c\epsilon_n$  for sufficiently large  $n$ .

## References

- AITKIN, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 111–142.
- AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, B. N. Petrov & F. Csaki, eds. Budapest: Akadémiai Kiado.
- AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.
- AKAIKE, H. (1985). Prediction and entropy. In *A celebration of statistics*. Springer, New York, pp. 1–24.
- BAHADUR, R. R. (1967). An optimal property of the likelihood ratio statistic. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press.
- BARRON, A., BIRGÉ, L. & MASSART, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113**, 301–413.
- BARRON, A. R. (1998). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics 6*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds. New York: Oxford, pp. 27–52.
- BERNARDO, J. M. & SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York, NY.
- BHATTACHARYA, A., PATI, D. & DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics* **42**, 352–381.
- BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability theory and related fields* **71**, 271–291.
- BOLTZMANN, L. (1878). Weitere Bemerkungen über einige Probleme der mechanischen Wärmetheorie. *Wiener Berichte* **78**, 7–46.
- BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80**, 580–598.
- BUNKE, O. & MILHAUD, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *The Annals of Statistics* **26**, 617–644.
- CAMPBELL, L. L. (1966). Exponential entropy as a measure of extent of a distribution. *Probability Theory and Related Fields* **5**, 217–225.
- CASELLA, G. & MORENO, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101**, 157–167.
- CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *The Annals of Statistics* **42**, 2058–2091.

- CLYDE, M. A. & IVERSEN, E. S. (2013). Bayesian model averaging in the M-open framework. In *Bayesian Theory and Applications*, P. Damien, P. Dellaportas, N. G. Polson & D. A. Stephens, eds. Oxford University Press, pp. 483–498.
- FAZEL, M., HINDI, H. & BOYD, S. P. (2003). Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *American Control Conference, 2003. Proceedings of the 2003*, vol. 3.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.
- GEISSER, S. & EDDY, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.
- GELFAND, A. E. & DEY, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 501–514.
- GUTIÉRREZ-PEÑA, E., RUEDA, R. & CONTRERAS-CRISTÁN, A. (2009). Objective parametric model selection procedures from a Bayesian nonparametric perspective. *Computational Statistics & Data Analysis* **53**, 4255–4265.
- GUTIÉRREZ-PEÑA, E. & WALKER, S. G. (2005). Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review* **73**, 309–330.
- HJORT, N. L., HOLMES, C., MÜLLER, P. & WALKER, S. G., eds. (2010). *Bayesian nonparametrics*. Cambridge University Press, Cambridge.
- HOFFMANN, M. & LEPSKI, O. (2002). Random rates in anisotropic regression. *The Annals of Statistics* **30**, 325–358.
- IKEDA, S. (1960). A remark on the convergence of Kullback-Leibler’s mean information. *Annals of the Institute of Statistical Mathematics* **12**, 81–88.
- JIANG, W. & TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics* **36**, 2207–2231.
- KASS, R. E. & WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934.
- KLEIJN, B. J. K. & VAN DER VAART, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics* **6**, 354–381.
- LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. & BERGER, J. O. (2008). Mixtures of  $g$  priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**, 410–423.
- OWHADI, H., SCOVEL, C. & SULLIVAN, T. (2015). On the brittleness of Bayesian inference. *SIAM Review* **57**, 566–582.
- RASMUSSEN, C. E. & WILLIAMS, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.
- VAN DER VAART, A. & VAN ZANTEN, H. (2011). Information rates of nonparametric Gaussian process methods. *The Journal of Machine Learning Research* **12**, 2095–2119.
- VAN DER VAART, A. & VAN ZANTEN, J. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, vol. 3. Beachwood, Ohio, USA: Institute of Mathematical Statistics, pp. 200–222.

- VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36**, 1435–1463.
- VAN DER VAART, A. W. & VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics* **37**, 2655–2675.
- WALKER, S., DAMIEN, P. & LENK, P. (2004). On priors with a Kullback–Leibler property. *Journal of the American Statistical Association* **99**, 404–408.
- YOO, W. W. & GHOSAL, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics* **44**, 1069–1102.



# Supplementary material for “A framework for probabilistic inferences from imperfect models”

Meng Li and David B. Dunson  
Department of Statistical Sciences, Duke University  
meng.li@stat.duke.edu dunson@duke.edu

November 3, 2016

## 1 Technical lemmas

The following Lemma 1.1 and Lemma 1.2 are used to convert the contraction rate of  $\widetilde{\text{KL}}_1(f_0, f_j)$  to  $\widetilde{\text{KL}}_2(f_0, f_j)$  in Theorem 1. Both lemmas are established under a general setting without requiring model (9) or Assumptions (a), (b) and (c). The proof of Lemma 1.1 uses the convexity of the Kullback-Leibler divergence and Jensen’s inequality. Lemma 1.2 is a generalized result of Lemma 8.2 in Kleijn & Van Der Vaart (2006) and we allow both arguments in  $\text{KL}(\cdot, \cdot)$  to depend on the sample size  $n$ .

**Lemma 1.1.** *For  $\widetilde{\text{KL}}_1(f_0, f_j)$  in (7) and  $\widetilde{\text{KL}}_2(f_0, f_j)$  in (8), the inequalities  $\widetilde{\text{KL}}_1(f_0, f_j) \geq \widetilde{\text{KL}}_2(f_0, f_j)$  hold for any sample size.*

*Proof of Lemma 1.1.* We first assert that for univariate densities  $g_1(x | w)$ ,  $g_2(x)$  and  $\pi(w)$ , we have

$$\int \text{KL}\{g_1(\cdot | w), g_2\} \pi(w) dw \geq \text{KL}(\widehat{g}_1, g_2), \quad \int \text{KL}\{g_2, g_1(\cdot | w)\} \pi(w) dw \geq \text{KL}(g_2, \widehat{g}_1),$$

where  $\widehat{g}_1(x) = \int g_1(x | w) \pi(w) dw$  and all  $\text{KL}(\cdot, \cdot)$  are well defined. The proof of this assertion is obtained by the convexity of the Kullback-Leibler divergence and Jensen’s inequality, which is detailed below. Applying Jensen’s inequality, we obtain that

$$\int g_1(x | w) \log g_1(x | w) \pi(w) dw \geq \widehat{g}_1(x) \log \widehat{g}_1(x),$$

and

$$\int -\log g_1(x | w) \pi(w) dw \geq -\log \int g_1(x | w) \pi(w) dw = -\log \widehat{g}_1(x).$$

for any  $x$ , since both functions  $s \mapsto s \log s$  and  $s \mapsto -\log s$  are convex. Therefore,

$$\begin{aligned} \int \text{KL}\{g_1(\cdot | w), g_2\} \pi(w) dw &= \int \int g_1(x | w) \log \frac{g_1(x | w)}{g_2(x)} \pi(w) dx dw \\ &= \int \int g_1(x | w) \log g_1(x | w) \pi(w) dx dw - \int \int g_1(x | w) \log g_2(x) \pi(w) dx dw \\ &= \int \int g_1(x | w) \log g_1(x | w) \pi(w) dx dw - \int \widehat{g}_1(x) \log g_2(x) dx \\ &\geq \widehat{g}_1(x) \log \widehat{g}_1(x) dx - \int \widehat{g}_1(x) \log \{g_2(x)\} dx = \text{KL}(\widehat{g}_1, g_2). \end{aligned}$$

Similarly,

$$\begin{aligned}
\int \text{KL}\{g_2, g_1(\cdot | w)\} \pi(w) dw &= \int \int g_2(x) \log \frac{g_2(x)}{g_1(x | w)} \pi(w) dx dw \\
&= \int \int g_2(x) \log\{g_2(x)\} \pi(w) dw dx - \int \int g_2(x) \log\{g_1(x | w)\} \pi(w) dw dx \\
&\geq \int g_2(x) \log\{g_2(x)\} dx - \int g_2(x) \log\{\widehat{g}_1(x)\} dx = \text{KL}(g_2, \widehat{g}_1).
\end{aligned}$$

Therefore, for any given parameter in  $f_0(\cdot | \theta_0)$  and  $f_j(\cdot | \theta_j)$  holding other parameters fixed, the value of Kullback-Leibler divergence decreases if that parameter is integrated out inside the operation  $\text{KL}(\cdot, \cdot)$  rather than outside. Since  $\widetilde{\text{KL}}_2(f_0, f_j)$  uses the posterior predictive densities for all parameters, it follows that  $\widetilde{\text{KL}}_1(f_0, f_j) \geq \widetilde{\text{KL}}_2(f_0, f_j)$  by applying the assertion iteratively over the parameter space.  $\square$

**Lemma 1.2.** *If  $p_n, q_n, p_\infty, q_\infty$  are probability densities such that  $p_n \rightarrow p_\infty$  and  $q_n \rightarrow q_\infty$  as  $n \rightarrow \infty$ , then  $\liminf \text{KL}(p_n, q_n) \geq \text{KL}(p_\infty, q_\infty)$ .*

*Proof of Lemma 1.2.* Let  $g_n = p_n \log(q_n/p_n) = g_n I(q_n/p_n > 1) + g_n I(q_n/p_n \leq 1)$  where  $I(\cdot)$  denotes the indicator function, and  $g_\infty = p_\infty \log(q_\infty/p_\infty)$ . The function  $g_n I(q_n/p_n > 1)$  is nonnegative and is bounded by  $q_n$ , since  $0 \leq (\log x) I(x > 1) \leq x$  if  $x > 0$ . Therefore,  $g_n I(q_n/p_n > 1)$  is uniformly integrable and thus  $\int g_n I(q_n/p_n > 1) \rightarrow \int g_\infty I(q_\infty/p_\infty > 1)$  as  $n \rightarrow \infty$  by the dominated convergence theorem. The function  $g_n I(q_n/p_n \leq 1) \leq 0$ , and an application of Fatou's lemma gives  $\limsup_{n \rightarrow \infty} \int g_n I(q_n/p_n \leq 1) \leq \int g_\infty I(q_\infty/p_\infty \leq 1)$ . Consequently, we obtain that  $\limsup_{n \rightarrow \infty} \int g_n \leq \int g_\infty$  and thus  $\liminf_{n \rightarrow \infty} \int -g_n \geq \int -g_\infty$ .  $\square$

The following Lemma (1.3) gives analytical forms of the divergence  $\delta_j$  for linear regression models when the true model is (9).

**Lemma 1.3** (Evaluation of  $\delta_j$ ). *Consider random designs where the covariate  $x$  is random and suppose Assumptions (a), (b) and (c) hold. and a linear model  $\mathcal{M}_j$  where the mean function  $\mu_j(x; \beta_j) = (1, x_j^T) \beta_j$  and  $x_j$  is a  $p_j$ -dimensional sub-vector of  $x$ . If  $\beta_j$  and  $\sigma_j$  follow prior distributions with support being  $\mathbb{R}^{p_j+1}$  and  $\mathbb{R}^+$ , then the divergence  $\delta_j$  is*

$$\delta_j = \frac{1}{2} \log \left[ 1 + \frac{\text{var}\{\mu(x)\} - \text{cov}\{x_j, \mu(x)\}^T \{\text{var}(x)\}^{-1} \text{cov}\{x_j, \mu(x)\}}{\sigma^2} \right]. \quad (1)$$

*Proof of Lemma 1.3.* The conditional Kullback-Leibler divergence between  $N\{\mu(x), \sigma^2\}$  and  $N\{\mu_j(x; \beta_j), \sigma_j^2\}$  given  $x$  is

$$\log \frac{\sigma_j}{\sigma} + \frac{\sigma^2 + \{\mu(x) - \mu_j(x; \beta_j)\}^2}{2\sigma_j^2} - \frac{1}{2}.$$

Applying the chain rule of the Kullback-Leibler divergence, we obtain the minimal Kullback-Leibler divergence  $\delta_j$  by taking the expectation with respect to  $x$ , namely,

$$\begin{aligned}
\delta_j &= \inf_{\theta_j \in \Theta_j} E \left[ \log \frac{\sigma_j}{\sigma} + \frac{\sigma^2 + \{\mu(x) - \mu_j(x; \beta_j)\}^2}{2\sigma_j^2} - \frac{1}{2} \right] \\
&= \inf_{\sigma_j \in \mathbb{R}} \left[ \log \frac{\sigma_j}{\sigma} + \frac{\sigma^2 + \inf_{\beta_j \in \mathbb{R}^{p_j+1}} E\{\mu(x) - \mu_j(x; \beta_j)\}^2}{2\sigma_j^2} - \frac{1}{2} \right].
\end{aligned}$$

For linear model  $\mathcal{M}_j$ , we have  $\mu_j(x) = (1, x^T) \beta_j = \beta_{j,1} + x^T \beta_{j,-1}$  where  $\beta_{j,1}$  is the intercept coefficient and  $\beta_{j,-1}$  is the slope coefficient. It is easy to see that the expectation  $E\{\mu(x) - \mu_j(x; \beta_j)\}^2$  is minimized when  $\beta_{j,1} = E\{\mu(x) - x^T \beta_{j,-1}\}$  at which  $E\{\mu(x) - \mu_j(x; \beta_j)\}^2 = \beta_{j,-1}^T \text{cov}(x, x) \beta_{j,-1} -$

$2\beta_{j,-1}^T \text{cov}\{x, \mu(x)\} + \text{var}\{\mu(x)\}$ . This is a quadratic form of  $\beta_{j,-1}$  achieving its minimum  $\text{var}\{\mu(x)\} - \text{cov}\{x_j, \mu(x)\}^T \{\text{var}(x)\}^{-1} \text{cov}\{x_j, \mu(x)\}$  at  $\beta_{j,-1} = \{\text{var}(x)\}^{-1} \text{cov}\{x, \mu(x)\}$ .

It is easy to see that

$$\inf_{\beta_j \in \mathbb{R}^{p+1}} E\{\mu(x) - \mu_j(x; \beta_j)\}^2 = \text{var}\{\mu(x)\} - \text{cov}\{x_j, \mu(x)\}^T \{\text{var}(x)\}^{-1} \text{cov}\{x_j, \mu(x)\}.$$

If  $\sigma_j$  follows a prior distribution with support being the positive line, then  $\delta_j$  is minimized when  $\sigma_j^2 = \sigma^2 + \inf_{\beta_j \in \mathbb{R}^{p+1}} E\{\mu(x) - \mu_j(x; \beta_j)\}^2$ , which concludes that

$$\delta_j = \frac{1}{2} \log \left[ 1 + \frac{\text{var}\{\mu(x)\} - \text{cov}\{x_j, \mu(x)\}^T \{\text{var}(x)\}^{-1} \text{cov}\{x_j, \mu(x)\}}{\sigma^2} \right].$$

This completes the proof.  $\square$

**Remark 1.4.** If the covariate  $x_j$  is univariate, the divergence  $\delta_j$  in (1) can be simplified as

$$\delta_j = \frac{1}{2} \log \left( 1 + \frac{\text{var}\{\mu(x)\}}{\sigma^2} [1 - \rho^2\{x_j, \mu(x)\}] \right),$$

where  $\rho\{x_j, \mu(x)\}$  is the correlation between  $x_j$  and  $\mu(x)$ .

## 2 Additional simulation results

In this section, we provide details of additional simulation results using the four cases in the simulation section. The four cases with different mean functions are given in Figure 1. Case 1 and Case 2 are for the  $\mathcal{M}$ -closed situation where the model list  $\mathcal{M}$  contains the true model, while Case 3 and Case 4 are for the  $\mathcal{M}$ -open situation. The oracle divergence  $\delta_j$  between model  $\mathcal{M}_j$  and the true model is reported in Table 1, which is calculated using Lemma 3 in the supplementary material. Table 1 also presents the estimates  $\widetilde{\text{KL}}_1(f_0, f_j)$  and  $\widetilde{\text{KL}}_2(f_0, f_j)$ . We can see that both  $\widetilde{\text{KL}}_t(f_0, f_j)$  quickly converge to the corresponding  $\delta_j$  in Case 1, 2 and 4. Case 3 corresponds to a subtle cyclic deviation from Case 2; we find in this case that the reference nonparametric model fails to pick up the cyclic deviation so that the estimates of  $\widetilde{\text{KL}}_t(f_0, f_j)$  are close to Case 2 but deviate from  $\delta_j$ . However, the estimates of  $\delta_N - \delta_F$  are accurate, suggesting robustness of the conditional D-probabilities to performance of the reference model.

Figure 3 plots histograms of D-probabilities  $\pi_{j,t}$  for all four cases when  $n = 100$ . We select the model with the maximum D-probability in each replication and calculate the model selection probability across 1000 replications. In Case 1, the D-probability is high for the true model  $\mathcal{M}_F$  in all replications, suggesting model selection uncertainty close to zero. Model  $\mathcal{M}_F$  has absolute D-probabilities that are not close to zero, such as 0.3, providing evidence it is an adequate approximation. In Case 2, both models have high D-probabilities as expected. The inclusion probability of the covariate  $x$  is either 0.09 using  $\pi_{j,1}$  or 0.23 using  $\pi_{j,2}$ , suggesting preference for the null model, with  $\pi_{j,1}$  providing a greater penalty on model complexity as in (14). The slight difference in scales between  $\pi_{j,1}$  and  $\pi_{j,2}$  is less prominent for conditional D-probabilities as shown in Figure 2. In Case 4, the full model  $\mathcal{M}_F$  is assigned probability 1, but the D-probabilities are both close to zero, suggesting lack of fit.

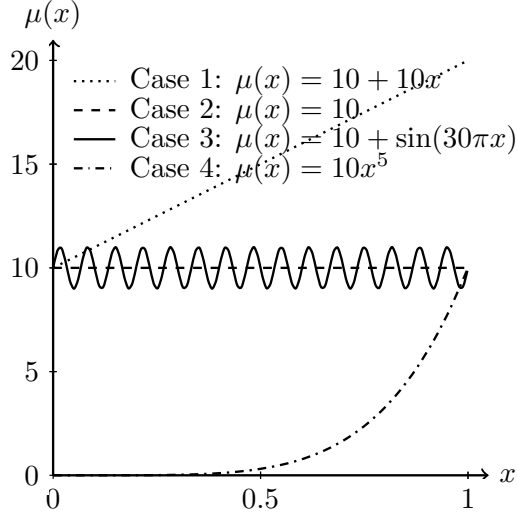


Figure 1: Mean functions for the four cases.

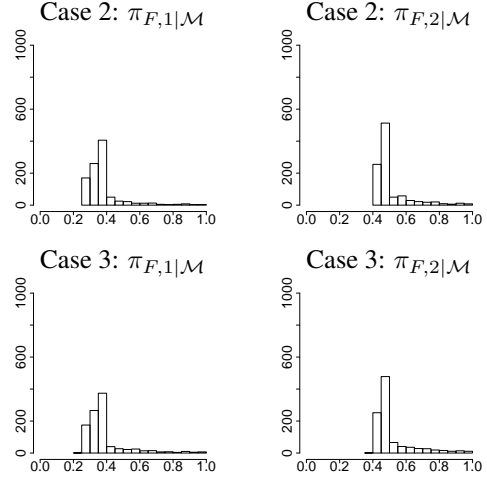


Figure 2: Conditional D-probabilities  $\pi_{F,1|\mathcal{M}}$  and  $\pi_{F,2|\mathcal{M}}$  of the full model in Case 2 and 3.

Table 1: Comparison of  $\widetilde{\text{KL}}_t(f_0, f_j)$  and  $\delta_j$  for  $t = 1, 2$ . In each case, we report the estimates for  $\mathcal{M}_F$ ,  $\mathcal{M}_N$  and their differences in the row  $N - F$ , averaged across 1000 replications. The last column reports the maximum standard errors of the estimates in each row. All estimates and standard errors are multiplied by 100.

(Case, $\mathcal{M}_j$ )	$\delta_j$	$\widetilde{\text{KL}}_1(f_0, f_j)$		$\widetilde{\text{KL}}_2(f_0, f_j)$		SE
Sample Size		100	500	100	500	
(1, $\mathcal{M}_F$ )	0	3.79	0.76	1.50	0.30	0.02
(1, $\mathcal{M}_N$ )	111.68	112.16	111.86	111.01	111.62	0.26
(1, $N - F$ )	111.68	108.37	111.10	109.51	111.32	0.26
(2, $\mathcal{M}_F$ )	0	2.94	0.58	1.28	0.25	0.02
(2, $\mathcal{M}_N$ )	0	2.44	0.48	1.34	0.26	0.03
(2, $N - F$ )	0	-0.5	-0.1	0.05	0.01	0.11
(3, $\mathcal{M}_F$ )	20.23	2.96	0.61	1.27	0.27	0.02
(3, $\mathcal{M}_N$ )	20.27	2.49	0.54	1.38	0.31	0.03
(3, $N - F$ )	0.05	-0.46	-0.07	0.11	0.05	0.26
(4, $\mathcal{M}_F$ )	55.94	56.35	55.93	53.87	55.41	0.27
(4, $\mathcal{M}_N$ )	99.48	99.28	99.41	97.23	98.98	0.36
(4, $N - F$ )	43.54	42.94	43.48	43.36	43.57	0.11

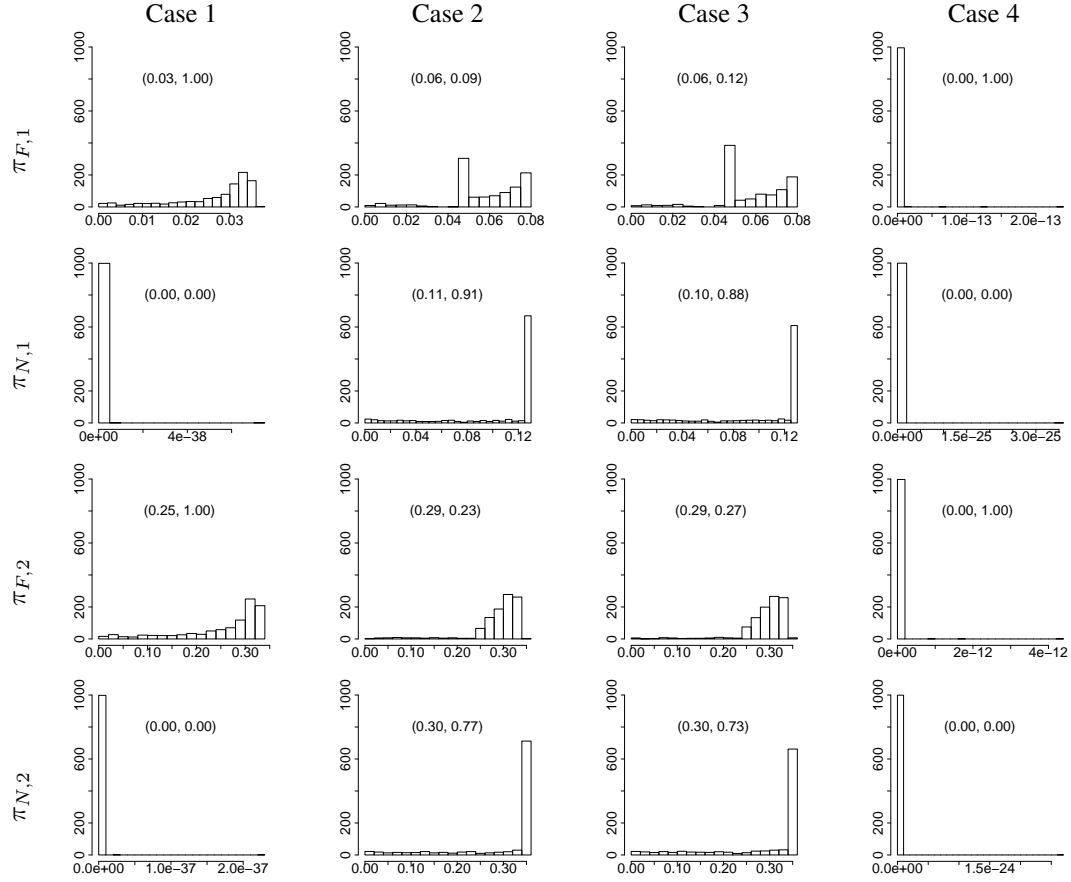


Figure 3: Histograms of D-probabilities for all four cases when  $n = 100$ . The two numbers  $(p_1, p_2)$  in the middle of each histogram are the average D-probability and the selection probability among 1000 replications.

### 3 Description of variables in Ozone data

ozone	Daily maximum of the hourly average ozone concentrations in Upland, CA
vh	500 millibar pressure height, measured at the Vandenberg air force base
wind	Wind speed in mph at LAX airport
humidity	Humidity in percent at LAX
temp	Sandburg Air Force Base temperature in degrees Fahrenheit
ibh	Temperature inversion base height in feet
dpg	Pressure gradient from LAX to Daggert in mm Hg
ibt	Inversion base temperature at LAX in degrees Fahrenheit
vis	Visibility at LAX in miles

### References

KLEIJN, B. J. K. & VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* **34**, 837–877.